

IN THE NAME OF GOD

AN INTELLIGENT AGENT-ORIENTED STRUCTURE FOR
TEXT INFORMATION RETRIEVAL

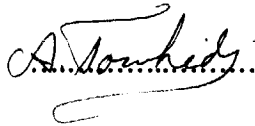
BY
MAZIAR SALEHI

THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN
PARTIAL FULFILLMANT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (M.Sc)

IN
COMPUTER ENGINEERING-
ARTIFICIAL INTELLIGENCE AND ROBOTICS
SHIRAZ UNIVERSITY
SHIRAZ, IRAN

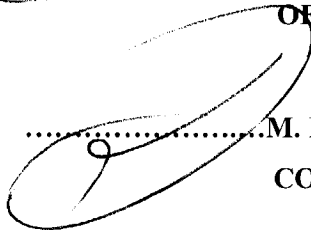
EVALUATED AND APPROVED BY THE THESIS COMMITTEE AS: EXCELLENT



.....A. TOWHIDI, Ph.D., ASSISTANT PROF. OF
COMPUTER ENGINEERING (CHAIRMAN)



.....M. ZOLGHADRI JAHROMI, Ph.D., ASSISTANT PROF.
OF COMPUTER ENGINEERING (CHAIRMAN)



.....M. H. SADREDDINI, Ph.D., ASSISTANT PROF. OF
COMPUTER ENGINEERING

November 2000

۲۳.۵۴

*In memory of my father,
The mount of patience and endurance;
And to my mother,
The foaming spring of affection and love;
And to all those whom I care for,
Without knowing it.*

Great discoveries and achievements invariably
involve the cooperation of many minds.

Alexander Graham Bell

I hear and I forget. I see and I believe. I do and I
understand.

Confucius

29.03

ACKNOWLEDGMENT

The author wishes to express the deepest and sincere thanks and appreciation to his supervisors *Dr. A. Towhidi* and *Dr. M. Zolghadri Jahromi* for their supervision, guidance, contributions and encouragement and for kindly and patiently giving many hours of their time.

The author would also like to express his gratitude to the member of the thesis committee, *Dr. M. H. Sadreddini* for his valuable suggestion and criticism. Thanks are also due to *Dr. Khodaparasti*, from department of foreign languages, for his guidance and cooperation.

Thanks are also due to all of my entire colleague students and friends whose constant encouragement have improved the quality of this work and have modestly requested not to be mentioned, but especially *Mr. A. Sajjadih* deserves to be mentioned here.

Finally, the author expresses his warmest and most passionate regards to his family who are the dearest moral supporters.

ABSTRACT

AN INTELLIGENT AGENT-ORIENTED STRUCTURE FOR TEXT INFORMATION RETRIEVAL

**BY
MAZIAR SALEHI**

As the volume of information available on the Internet and corporate intranets is increasing, there is a growing need for tools to help people better find, filter and manage these resources. Artificial intelligence methods, by their adaptable and flexible properties, can help us handle the complexity of these new tools.

This thesis deals with modeling information retrieval (IR) systems using an agent-oriented (AO) approach. The AO approach, which is a new paradigm in software engineering, tries to model software modules as autonomous, intelligent creatures, which cooperate with each other to achieve the desired goal(s). In this work, based on modeling techniques proposed by other researchers, a new agent-oriented technique is proposed, and a prototype system is designed using this method.

In the field of text retrieval, different retrieval methods are investigated and their performances are evaluated using standard test collections (CRANFIELD, MEDLINE). Latent Semantic

Indexing technique is used in this work to extract the associative relationship between terms and documents. A genetic algorithm is also developed to enhance the performance of the matching between queries and documents.

TABLE OF CONTENTS

CONTENT	PAGE
LIST OF TABLES.....	X
LIST OF FIGURES.....	XII
CHAPTER I: INTRODUCTION.....	1
CHAPTER II: INFORMATION RETRIEVAL.....	5
2.1 Introduction	5
2.2 What is IR?	5
2.2.1 Functional approach	6
2.2.2 Basic Information Retrieval system	7
2.3 Document Indexing	14
2.3.1 File format conversion.....	15
2.3.2 Text segmentation	15
2.3.3 Term Extraction	16
2.3.4 Dimensionality control	17
2.3.4.1 Stop word elimination.....	18
2.3.4.2 Word Stemming	18
2.4 Retrieval models	20
2.4.1 Statistical Models	20
2.4.1.1 Vector space model	21
2.4.1.2 Term Weighting	21
2.4.1.3 Similarity function.....	29
2.4.1.4 Dimensionality reduction.....	31
2.4.1.5 Concept space model	34
2.4.2 Probabilistic model.....	35
2.4.3 Knowledge based model.....	35
2.4.4 Models based on machine learning.....	36
2.5 Retrieval Evaluation.....	37
2.5.1 Evaluation of effectiveness.....	38
2.5.1.1 Evaluation viewpoints and the relevance problem.....	38
2.5.2 Measures of retrieval effectiveness.....	39
2.5.2.1 Recall, Precision and fallout	39
2.5.2.2 Single-Valued measures.....	43

CONTENT	PAGE
CHAPTER III: INTELLIGENT SOFTWARE AGENTS	45
3.1 Introduction	45
3.2 Theory of agency	46
3.2.1 What is an agent?	46
3.2.2 The Essence of Agency	53
3.2.3 Agent Classification	57
3.2.4 What Agents are not	63
3.3 Agent view from software engineering	64
3.3.1 Complexity in software engineering	64
3.3.1.1 Canonical view of AOP	67
3.4 OO modeling	70
3.4.1 Different methodologies	70
3.4.2 UML	71
3.5 Agent-oriented modeling	72
3.5.1 Agents versus objects	72
3.5.1.1 AOP/OOP Comparison	74
3.6 Agent in Practice	75
3.6.1 Applications	75
3.6.1.1 Industrial Applications	75
3.6.1.2 Commercial Applications	77
3.6.1.3 Medical Applications	79
3.6.1.4 Entertainment	80
3.6.2 Languages	80
3.6.2.1 Java	81
3.6.2.2 Agent-0, PLACA, Agent-K	82
3.6.2.3 Discussion	86
CHAPTER IV: LATENT SEMANTIC INDEXING	89
4.1 Introduction	89
4.2 Problems in indexing large text collections	90
4.3 The QR factorization	95
4.3.1.1 Identifying a basis for the column space	96
4.3.1.2 The geometry of the vector space model	98
4.3.1.3 The low-rank approximation	100
4.4 The Singular Value Decomposition	102
4.4.1 Choosing the number of factors	105
4.4.2 Queries	107
4.4.3 Updating	107
4.5 Other methods for LSI	110

CONTENT	PAGE
4.5.1 Random projection	110
4.5.2 Semi-Discrete matrix Decomposition	111
4.6 Applications of LSI.....	112
4.6.1 Information Retrieval	112
4.6.2 Relevance Feedback	113
4.6.3 Information Filtering	114
4.6.4 TREC.....	115
4.6.5 Novel applications.....	117
4.6.5.1 Cross-Language Retrieval.....	118
4.6.5.2 Modeling Human Memory.....	120
4.6.5.3 Matching People instead of Documents	121
4.6.5.4 Noisy Input.....	122
CHAPTER V: EXPERIMENTS AND RESULTS.....	124
5.1 Introduction	124
5.2 Indexing.....	124
5.2.1 Feature extraction.....	125
5.2.2 Stemming.....	126
5.3 Term weighting.....	127
5.4 Choosing the results	133
5.5 Retrieval evaluation	133
5.6 Normalization effect.....	136
5.7 Similarity measurement.....	140
5.8 Latent semantic indexing.....	146
5.8.1 Elicitation of Semantics.....	146
5.8.2 Choosing the reduction factor.....	154
5.8.3 LSI Effect on retrieval	155
5.9 Relevance feedback.....	158
CHAPTER VI: MODELING TEXT RETRIEVAL SYSTEM.....	164
6.1 Introduction	164
6.2 AO modeling	166
6.2.1 Several proposed models	166
6.2.2 Modeling method	168
6.2.2.1 Problem domain analysis with UML.....	169
6.2.2.2 Agent elicitation	170
6.2.2.3 Intra-agent modeling.....	171
6.2.2.4 Inter-agent Modeling	172
6.3 Analysis the system.....	173

CONTENT	PAGE
6.3.1 Problem definition.....	174
6.3.2 Decomposition.....	175
6.3.3 Scenarios.....	177
6.3.4 UML analysis.....	178
6.3.4.1 Use cases.....	178
6.3.4.2 Sequence diagrams.....	181
6.4 Modeling the agent-based system.....	184
6.4.1 Incremental development.....	184
6.4.2 Agent-Class diagram.....	184
6.4.3 Inter-Agent modeling.....	185
6.4.3.1 Belief modeling.....	185
6.4.3.2 Goal modeling.....	186
6.4.3.3 Plan modeling.....	187
6.4.3.4 Capabilities.....	189
6.4.4 Intra-Agent modeling.....	190
6.4.4.1 Agent communication modeling.....	191
6.5 Infrastructure.....	191
6.5.1 Agent Platform.....	192
6.5.2 Agent Management System.....	194
6.5.3 Directory Facilitator.....	195
6.5.4 Agent Platform Security Manager.....	195
6.5.5 Agent Communication Channel.....	196
CHAPTER VII: CONCLUSIONS AND FUTURE WORKS.....	197
7.1 Conclusions.....	197
7.1.1 Information retrieval.....	197
7.1.2 Agent-based modeling of the system.....	200
7.2 Future works.....	203
Appendix A: UML tables and use cases of system.....	205
REFERENCES	
ABSTRACT AND TITLE PAGE IN PERSIAN	

LIST OF TABLES

TABLE	PAGE
Table 2-1 Sample list of stop words. [2].....	18
Table 2-2 Typical suffix list [2].....	19
Table 2-3 Established local weight formulas used.	22
Table 2-4 Established global weight formulas used.....	26
Table 2-5 Normalization factors used.....	28
Table 2-6 Popular weighting schemes.....	29
Table 3-1 Comparison of agent and object-oriented programming	74
Table 5-1 Characteristics of test collections	124
Table 5-2 Additional Local weighting formulas	128
Table 5-3 Additional Global weightings.....	128
Table 5-4 Improvement versus Raw TF weighting (CRAN).....	131
Table 5-5 Improvement versus Raw TF (MED).....	133
Table 5-6 original Term-Doc matrix	148
Table 5-7 U matrix in SVD.....	148
Table 5-8 Singular value matrix.....	149
Table 5-9 V matrix in SVD decomposition	149
Table 5-10 Reconstructed matrix using k=2	149
Table 5-11 Co-relation matrix in original term-document matrix.....	151
Table 5-12 Co-relation matrix in new term-document matrix	152
Table 5-13 LSI retrieval improvement versus another weighting (MED). 157	
Table 5-14 Relevance feedback with the first Rel. Doc (MED)	161
Table 5-15 Relevance feedback with the first three Rel. Docs (MED).....	161
Table 5-16 Relevance feedback with all Rel. Docs (MED).....	161
Table 5-17 Relevance feedback with the first Rel. Doc (MED with LSI). 162	

TABLE	PAGE
Table 5-19 Relevance feedback with all Rel. Doc (MED with LSI).....	163
Table 6-1 Evolution of programming approaches.....	165
Table 6-2 Agent selection rules.....	171
Table 6-3 Use case general template	179
Table 6-4 Agents' general goals.....	186

LIST OF FIGURES

FIGURE	PAGE
Figure 2-1 Information system environment [2]	7
Figure 2-2 Functional overview of information retrieval [2].....	7
Figure 2-3 Basic Information retrieval system architecture.....	8
Figure 2-4 Information organization	10
Figure 2-5 Indexing process.....	11
Figure 2-6 Text retrieval process.....	12
Figure 2-7 Query processing.....	13
Figure 2-8 Document searching	13
Figure 2-9 Showing relevant documents	14
Figure 2-10 Cosine similarity measurement.	31
Figure 2-11 Histogram of word frequencies showing Zipf's Law.....	33
Figure 2-12 partitioning text collection versus query.....	40
Figure 2-13 typical average recall-precision graph.....	41
Figure 3-1 Agents and their environments [31]	49
Figure 3-2 A Part View of an Agent Typology.....	59
Figure 3-3 View of a Canonical Complex System.....	66
Figure 3-4 Canonical view of a multi-agent system.....	68
Figure 4-1 Mathematical representation of the matrix A_k	104
Figure 4-2 Mathematical representation of folding-in p documents.	109
Figure 4-3 Mathematical representation of folding-in q terms.	110
Figure 5-1 Comparison of weightings (CRAN 100 docs, 50 queries)	130
Figure 5-2 comparison of weightings (CRAN 1400 docs 225 queries)	131
Figure 5-3 Comparison of several basis weightings (MED 1033 docs, 30 query)	132

FIGURE	PAGE
Figure 5-4 Comparison of new weightings versus previous ones	132
Figure 5-5 Normalization effect	140
Figure 5-6 Evolution of matching function by E as fitness (CRAN sub.) ..	144
Figure 5-7 Evolution of matching function by DC_ARP as fitness (CRAN sub.).....	144
Figure 5-8 Evolution of matching function by IAP as fitness (CRAN sub.)	145
Figure 5-9 performance versus number of dimension (MED).....	156
Figure 5-10 Comparison between previous weightings and LSI.....	157
Figure 5-11 Positive and negative feedback	160
Figure 5-12 Relevance feedback with all Rel. Docs (MED)	162
Figure 5-13 Relevance feedback with all Rel. Doc (MED with LSI)	163
Figure 6-1 internal model of Agent based on BDI.....	168
Figure 6-2 Agent-oriented software development process mode.....	169
Figure 6-3 Agent-oriented modeling process.....	170
Figure 6-4 Information system general overview	174
Figure 6-5 Initial system decomposition.....	175
Figure 6-6 Final System decomposition	176
Figure 6-7 Middle system use case diagram	179
Figure 6-8 Interface system use case diagram	180
Figure 6-9 Retrieval system use case diagram	180
Figure 6-10 Middle system collaboration diagram.....	181
Figure 6-11 Interface system collaboration diagram.....	182
Figure 6-12 Retrieval system collaboration diagram (1)	183
Figure 6-13 Retrieval system collaboration diagram (2)	183
Figure 6-14 Agent-Class diagram	185
Figure 6-15 System goal hierarchy.....	187

FIGURE	PAGE
Figure 6-16 Plan model of interface agent.....	187
Figure 6-17 Plan model of Retrieval Agent.....	188
Figure 6-18 plan model of Middle Agent.....	188
Figure 6-19 Capability diagram of interface Agent.....	189
Figure 6-20 Capability diagram of Middle Agent.....	189
Figure 6-21 Capability diagram of Retrieval Agent.....	190
Figure 6-22 communication model of Agents.....	191
Figure 6-23 JATLite architecture.....	192
Figure 6-24 Agent management reference model.....	193
Figure 6-25 Agent platform architecture.....	194
Figure 7-1 Broker system.....	202
Figure 7-2 MatchMaker system.....	202

CHAPTER I

INTRODUCTION

The growth of the Internet and the availability of enormous volumes of data (information explosion) have necessitated intense interest in techniques to assist the user in locating data of interest. Internet has covered over 350 million pages of data till 1998 [1] and is predicted to reach over several billions in first decade of 21st century. The digital library effort is also progressing with the goal of migrating from the traditional book environment to a digital library environment.

Information retrieval proposes methods for organizing and modeling information and user's needs. IR methods basically are founded on information theory and they try to automatically perform all of the retrieval processes. Initial works are performed on text items (news, reports...) but in later years other information items (image, audio, etc) are also involved. However, most of information still consists of huge amount of text items, so text retrieval approaches are still as important portion of IR.

Several models are proposed for intelligent IR with the aid of statistics, probability theory and machine learning techniques. Statistical model such as vector space, which empowered by information theory, is one of the initial successful models.